* RELATÓRIO TÉCNICO *

O SUBSISTEMA DE MEMÓRIA DE MASSA DO MULTIPROCESSADOR MULTIPLUS

Sidney de Castro Oliveira
Júlio Salek Aude

NCE 34/90

Universidade Federal do Rio de Janeiro
Núcleo de Computação Eletrônica
Caixa Postal 2324
20001 - Rio de Janeiro - RJ
BRASIL

* Este artigo foi submetido originalmente nos Anais do III Simpósio Brasileiro de Arquitetura de Computadores e Processamento Paralelo (SBAC-PP), novembro de 1990, Rio de Janeiro

UNIVERSIDADE FEDERAL DO RIO DE JANEIR
NUCLEO DE COMPUTACAO ELETRONICA

O Subsistema de Memória de Massa do Multiprocessador MULTIPLUS

RESUMO

Este artigo descreve a arquitetura de E/S de memória de massa proposta para o MULTIPLUS, um multiprocessador científico de alto desempenho em desenvolvimento no NCE/UFRJ. Após caracterizar o gargalo de E/S, decorrente de um descompasso existente na evolução tecnológica de microprocessadores e dos dispositivos de memória de massa, e de identificar os parâmetros que determinam o desempenho de um subsistema de E/S, o artigo discute as vantagens e desvantagens de um subsistema de E/S concentrado ou distribuido. Finalmente, uma arquitetura distribuida, capaz de facilitar a exploração de paralelismo em aplicações científicas, é proposta para o subsistema de E/S do MULTIPLUS e a organização interna dos processadores de E/S é apresentada e justificada através de uma análise quantitativa.

The Mass Storage Subsystem of MULTIPLUS Multiprocessor

ABSTRACT

This paper presents the architecture of the I/O subsystem for mass storage which has been proposed for MULTIPLUS, a high performance scientific multiprocessor under development at NCE/UFRJ. After studying the I/O bottleneck problem, which results from an existing gap in the technological evolution of processor and mass storage devices, and identifying the parameters which define the I/O subsystem performance, the paper discusses the advantages and drawbacks of the use of a concentrated or distributed I/O subsystem architecture. Finally, a distributed architecture, which is suitable for scientific parallel application, is proposed for the MULTIPLUS I/O subsystem and the internal organization of the I/O Processor is presented and justfied through numerical analysis.

O Subsistema de Memória de Massa do Multiprocessador MULTIPLUS

Sidney de Castro Oliveira
NCE/UFRJ

Júlio Salek Aude
NCE/IM/UFRJ

RESUMO

Este artigo descreve a arquitetura de E/S de memória de massa proposta para o MULTIPLUS, um multiprocessador científico de alto desempenho em desenvolvimento no NCE/UFRJ. Após caracterizar o gargalo de E/S, decorrente de um descompasso existente na evolução tecnológica de microprocessadores e dos dispositivos de memória de massa, e de identificar os parâmetros que determinam o desempenho de um subsistema de E/S, o artigo discute as vantagens e desvantagens de um subsistema de E/S concentrado ou distribuido. Finalmente, uma arquitetura distribuida, capaz de facilitar a exploração de paralelismo em aplicações científicas, é proposta para o subsistema de E/S do MULTIPLUS e a organização interna dos processadores de E/S é apresentada e justificada através de uma análise quantitativa.

ABSTRACT

This paper presents the architecture of the I/O subsystem for mass storage which has been proposed for MULTIPLUS, a high performance scientific multiprocessor under development at NCE/UFRJ. After studying the I/O bottleneck problem, which results from an existing gap in the technological evolution of processor and mass storage devices, and identifying the parameters which define the I/O subsystem performance, the paper discusses the advantages and drawbacks of the use of a concentrated or distributed I/O subsystem architecture. Finally, a distributed architecture, which is suitable for scientific parallel application, is proposed for the MULTIPLUS I/O subsystem and the internal organization of the I/O Processor is presented and justfied through numerical analysis.

Afiliação: Núcleo de Computação Eletrônica/UFRJ
Caixa Postal 2324
20001 - Rio de Janeiro, RJ
Tel: (021) 290 3212 R.328

S. C. Oliveira - Analista de Sistemas do NCE/UFRJ, Engenheiro Eletrônico pela UFRJ (1986). Áreas de Interesse: Arquiteturas de Computadores, Processamento Paralelo e Estações Gráficas de Trabalho.

J. S. Aude - Analista de Sistema do NCE/UFRJ, Professor Adjunto do Instituto de Matemática da UFRJ, Ph.D em Computação pela Universidade de Manchester (1986). Áreas de Interesse: CAD para VLSI, Arquitetura de Computadores e Processamento Paralelo.

1 - INTRODUÇÃO

O avanço da tecnologia de microeletrônica nos últimos anos tem resultado em dispositivos eletrônicos cada vez mais velozes e poderosos. A area de microprocessadores é uma das que mais tem se beneficiado com este avanço. Atualmente, a capacidade de processamento de informação dos microprocessadores é bastante elevada e crescente a cada ano. A utilização destes microprocessadores, principalmente em arquiteturas paralelas, tem aumentado muito a demanda por instruções e dados. Estas instruções e dados estão sempre presentes na memória do sistema, seja na memória principal, que é menor e mais rápida, seja na memória secundária, que são os dispositivos de armazenamento de massa (discos magnéticos, fitas, etc). Entretanto, enquanto a tecnologia de microeletrônica avança, proporcionando microprocessadores mais velozes e dispositivos de memória cada vez mais densos e baratos, a tecnologia dos dispositivos de armazenamento de massa não progride em igual proporção, criando um desequilíbrio entre a capacidade de processamento das máquinas e sua demanda por entrada e saída de dados. A minimização deste desequilíbrio, através de novas arquiteturas, tem sido bastante investigada pelos projetistas de computadores de alto desempenho.

Este trabalho resulta de um estudo de arquiteturas de entrada e saída de armazenamento de massa para computadores de alto desempenho, com o objetivo de definir um subsistema de E/S que atenda às necessidades do Multiprocessador Multiplus em desenvolvimento no NCE/UFRJ. A seção 2 caracteriza o gargalo do processamento de E/S, comparando a evolução tecnológica dos três componentes básicos de um sistema computacional: o microprocessador, a memória principal e a memória secundária. A seção 3 avalia o desempenho de um subsistema de E/S em função tanto do tipo de aplicação a que ele se destina, quanto da própria característica dos dispositivos de armazenamento. As opções para configuração de um subsistema de E/S é o tema da seção 4. A seção 5 expõe a arquitetura do subsistema de E/S adotado no Multiplus. Na seção 6, a organização interna do Processador de Entrada e Saída é apresentada juntamente com uma análise quantitativa que justifica a sua definição.

2 - CARACTERIZAÇÃO DO GARGALO DO PROCESSAMENTO DE E/S

Para melhor entender o desenvolvimento tecnológico que determina hoje a *performance* da CPU, da memória e dos dispositivos de entrada e saída, que resultou no desequilíbrio entre a capacidade de processamento da CPU e a dos dispositivos de E/S, será feita uma breve discussão sobre a evolução tecnológica de cada um deles nos últimos anos. Os microprocessadores, sem dúvida, foram os que mais se desenvolveram. Hoje é possível comprar por menos de mil dólares uma unidade do microprocessador Intel *i*860, cuja capacidade de processamento é similar ao do primeiro computador da linha CRAY (aproximadamente 33 Mips escalar). Usando a família de microprocessadores Intel pode-se observar que a capacidade de processamento vem dobrando a cada 2,25 anos. Considerando o surgimento comercial dos microprocessadores RISC em 1984, pode-se observar uma taxa de crescimento ainda maior, chegando a dobrar a capacidade de processamento a cada ano [HWANG 87].

Para suportar este aumento de demanda por instruções e dados

ocasionado pelo avanço tecnológico dos microprocessadores, o sistema de memória teve que se tornar maior e mais rápido. De forma aproximada, pode-se dizer que cada instrução por segundo do microprocessador requer um *byte* da memória principal, ou seja, 1M*Byte* para cada Mips. Isto sugere que a capacidade dos *chips* de memória acompanhe a taxa de crescimento da velocidade dos microprocessadores. Similarmente, fazendo-se uma análise da evolução da capacidade de armazenamento dos *chips* de memória no decorrer dos anos, pode-se observar que a capacidade de armazenamento vem aproximadamente quadruplicando a cada 3 anos nos últimos 20 anos. Entretanto, devido a rápida queda nos preços dos *chips* de memória, observa-se que a capacidade da memória principal tem crescido a uma taxa superior a da velocidade dos microprocessadores, chegando a 3 M*Bytes* por Mips em alguns sistemas.

Mas nem todas as instruções e dados requeridos pelo microprocessador estão presentes na memória principal do sistema. Muitas vezes é necessário buscar estas informações na memória secundária. Para manter todo o sistema em equilíbrio, é fundamental que o desempenho da memória secundária acompanhe o aumento de *performance* das outras partes do sistema. O principal elemento da memória secundária é o disco magnético. Para se medir o avanço tecnológico deste componente, considera-se a quantidade de *bits* armazenados por polegada quadrada, isto é, o número de *bits* por polegada em uma trilha vezes o número de trilhas por polegada. Pode-se observar que esta evolução tem permitido dobrar a capacidade de armazenamento a cada 3 anos. E também a cada 3 anos o preço destes discos reduz-se à metade [HWANG 87].

Para comparar o desenvolvimento tecnológico entre os microprocessadores, a memória principal e a memória secundária pode-se considerar que a velocidade dos microprocessadores vem dobrando a cada ano, que a densidade dos *chips* de memória vem dobrando a cada dois anos, e que a capacidade dos discos magnéticos vem dobrando a cada três anos. A Figura 1 ilustra graficamente esta situação. Pode-se notar que os últimos dez anos são suficientes para ocasionar uma grande discrepância entre o desenvolvimento tecnológico dos microprocessadores e das memórias, tanto em relação à memória principal quanto à memória secundária, principalmente. E pela situação atual da tecnologia, esta discrepância tende a aumentar, não havendo motivos para se acreditar na reversão desta tendência.

Por outro lado, a capacidade, tanto da memória principal quanto da memória secundária não é a única característica que precisa acompanhar o desenvolvimento tecnológico dos microprocessadores para manter o equilíbrio do sistema. Mesmo porque, a quantidade de memória é um fator que sempre pode crescer pela simples interligação de mais dispositivos de memória. Um outro fator até mais importante é a velocidade com que as instruções e dados requisitados pelos microprocessadores chegam até eles, ou seja, quão rápido os dispositivos de memória conseguem responder a um pedido de informação. No caso da memória principal, por dois motivos, ela vem acompanhando a velocidade dos microprocessadores. Primeiramente pela utilização de técnica de *caching*, onde um pequeno *buffer*, porém muito rápido, é colocado entre o microprocessador e a memória principal. Isto permite, pelas próprias características dos programas computacionais, que uma grande parte das informações requeridas pelos microprocessadores já estejam presentes neste *buffer*, possibilitando uma resposta muito mais rápida. O segundo motivo decorre do fato deste *buffer* ser implementado com memórias RAM estáticas, cuja velocidade tem dobrado a cada dois anos nos últimos anos. Desta forma, pode-se construir sistemas de memória principal extensos e muito rápidos (dependendo da taxa de acerto no *buffer*), capazes

de fornecer as informações requisitadas na velocidade máxima dos microprocessadores.

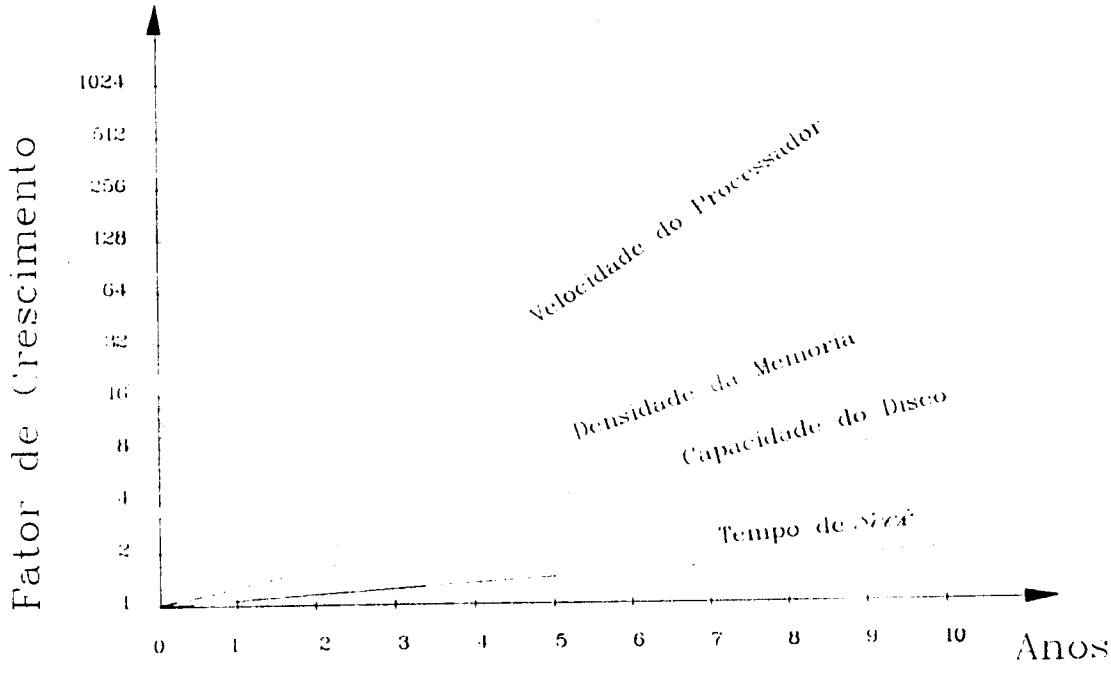

Figura 1: Comparação entre o desenvolvimento tecnológico dos principais elementos de um sistema computacional.

Para a memória secundária a situação é bem diferente. A *performance* dos discos magnéticos cresceu apenas modestamente. Os principais elementos que medem a velocidade dos discos são: o tempo de *seek* e a latência rotacional. O tempo de *seek* tem diminuido muito lentamente no decorrer dos anos, da ordem de 7% ao ano [HWANG 87]. Já a latência rotacional é proporcional a velocidade angular do disco, isto é, ao seu número de rotações por minuto. Entretanto, esta velocidade não se alterou, mantendo constante a latência rotacional ao longo dos anos.

Assim, pode-se verificar que o desenvolvimento tecnológico no setor de informática ocorreu de forma desigual entre os principais elementos de um computador: o microprocessador, a memória principal e a memória secundária. Basicamente, houve um progresso muito acentuado na área de microeletrônica, possibilitando o surgimento de *chips* muito mais poderosos, determinando a *performance* dos microprocessadores e da memória principal. Como a evolução da memória secundária dependeu de um progresso eletromecânico de seus dispositivos ou do surgimento de novos dispositivos de armazenamento, fatos que não ocorreram, surgiu um *gap* entre a *performance* da memória secundária e a dos microprocessadores. A busca de soluções para um maior equilíbrio entre estes elementos é um dos desafios atuais dos projetistas de arquiteturas de alto desempenho.

3 -CARACTERIZAÇÃO DO DESEMPENHO DOS SUBSISTEMAS DE E/S

O desempenho dos subsistemas de E/S está fortemente ligado ao tipo de processamento de entrada e saída associado aos subsistemas. Este processamento é função do tipo de aplicação a que o subsistema se destina. Antes de caracterizar o desempenho destes subsistemas é necessário esclarecer melhor alguns pontos sobre discos magnéticos e os elementos que determinam sua *performance*. Uma unidade de disco magnético é composta de um conjunto de *platter*, que consistem de discos metálicos cobertos por um material magnético onde são gravadas as informações. Cada *platter* contém uma quantidade de trilhas circulares. Estas trilhas são divididas em setores, que constituem, fisicamente, a menor quantidade de dados lidos ou escritos na unidade. Entretanto, as informações são armazenadas em blocos no disco, sendo que cada bloco consiste de uma quantidade fixa e definida de setores. As informações gravadas são recuperadas por uma cabeça de leitura/escrita posicionada em um braço que se move ao longo de cada *platter* através de um acionador.

A *performance* desta unidade está associada ao tempo gasto por ela para fornecer um bloco de informações dada a sua solicitação, e pode ser dividida em três componentes: o tempo de *seek*, a latência rotacional e o tempo de tranferência dos dados. O tempo de *seek* é o tempo necessário para o disco mover a cabeça de leitura/escrita até a trilha apropriada que contém o dado. Este tempo está associado a uma inércia inicial para tirar a cabeça da posição de repouso, que é da ordem de alguns mili-segundos, e ao número de trilhas a serem avançadas. Após a acelereção da cabeça, o tempo de avanço de trilha é bem inferior, chegando a um terço do inicial. Tipicamente, o tempo médio de *seek*, dadas duas trilhas aleatórias, se situa na faixa de 10 a 20ms.

O segundo componente é a latência rotacional, que é o tempo gasto para o setor de início de bloco dentro da trilha se posicionar tangencialmente à cabeça de leitura/escrita para permitir a leitura ou escrita dos dados. Este tempo é função do número de rotações por segundo do disco, que é atualmente de 3600 rpm. Desta forma, o tempo de uma revolução completa é de aproximadamente 16ms, e o tempo médio de latência, que é igual a metade do tempo de uma revolução, é de cerca de 8ms. Pode-se notar que, no pior caso, o tempo de latência é comparável ao tempo médio de *seek*.

O último componente é o tempo de transferência dos dados, ou seja, o tempo gasto para os *bytes* lidos serem transferidos do disco para o subsistema de entrada e saída ou vice-versa. Em contra-partida com o tempo de *seek* e de latência rotacional, que como visto anteriomente independem do tamanho do bloco, o tempo de transferência cresce com o tamanho do bloco. Assim sendo, existe um compromisso entre a definição do tamanho do bloco e o tipo de transferência. Se as transferências predominantes são de arquivos extensos, é conveniente escolher um tamanho de bloco grande, para o tempo de *seek* e de latência serem atenuados frente ao tempo de transferência dos dados. Caso contrário, se as transferências predominantes são de arquivos curtos, é conveniente escolher um tamanho de bloco pequeno, evitando transferir informações desnecessárias e otimizando a ocupação do disco.

O predomínio de transferências de arquivos extensos ou curtos está relacionado com a característica do processamento de entrada e saída referente ao tipo de aplicação a que a máquina se destina. Computadores de uso geral processam simultaneamente um grande número de pequenas tarefas,

que manipulam, cada uma, pequenas quantidades de dados. Por outro lado os computadores científicos processam poucas tarefas, mas que demandam grandes transferências de dados. A Figura 2 ilustra esta situação, onde é mostrada a composição dos três elementos que determinam a *performance* dos discos magnéticos em função do tipo de aplicação. Pode-se notar que o processamento geral gasta a maior parte do tempo de entrada e saída para movimentar a cabeça de leitura/escrita (tempo de *seek*) e na latência rotacional, de modo que qualquer avanço na taxa de transferência dos dados não traz grandes benefícios. Por outro lado, a distribuição do tempo de E/S nas aplicações científicas é mais equalitária entre o tempo de *seek* e de transferência de dados. Desta forma, sua *performance* é bastante sensível a qualquer avanço tecnológico do disco.

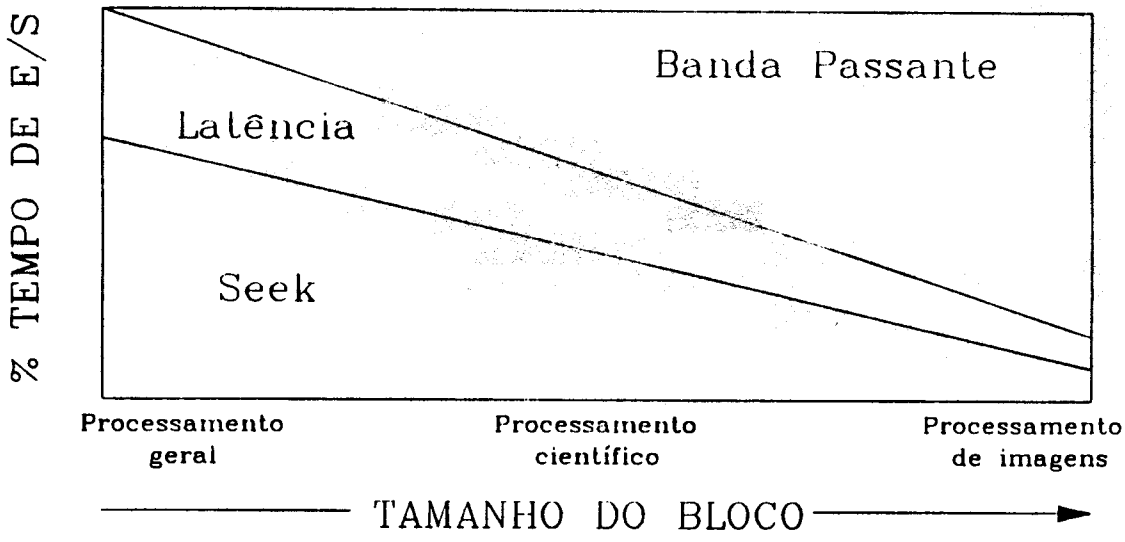

Figura 2: Distribuição do tempo gasto com E/S em função do tipo de aplicação submetida ao Processador de E/S.

Porém, a *performance* de um subsistema de entrada e saída não depende exclusivamente da *performance* do disco magnético, mas também da *performance* de todo o controlador de E/S. De uma maneira geral, um subsistema de E/S é composto de um processador de E/S, que controla todo o subsistema, um conjunto de dispositivos de armazenamento tais como discos magnéticos, e um canal de comunicação entre eles, permitindo a transferência dos dados. Analogamente aos discos magnéticos, a *performance* deste subsistema quando submetido a uma carga de processamento de E/S pode ser analisada em função de três fatores: *throughput*, latência e banda passante. *Throughput* refere-se ao número de pedidos de E/S atendidos por unidade de tempo. Latência representa o tempo gasto para um determinado pedido ser atendido, também conhecido como *overhead*. A banda passante avalia a quantidade de dados por unidade de tempo que flui entre os dispositivos de armazenamento e o controlador de E/S.

Baseado nisto, e em analogia com os discos magnéticos, pode-se também definir as características necessárias a um subsistema de E/S em função do tipo de aplicação a que ele se destina. A computação científica pode ser caracterizada, quase inteiramente, por operações de E/S

sequenciais. Tipicamente, os dados são transferidos em grande quantidade do disco para a memória principal, processados, e os resultados são periodicamente reescritos no disco. Este tipo de aplicação exige uma larga banda passante, com o mínimo de latência do controlador, para as transferências serem rápidas, já que são extensas. Entretanto, é caracterizado por um baixo *throughput*, pois são poucos os pedidos de E/S por unidade de tempo. Por outro lado, as operações de E/S nas aplicações de uso geral são caracterizadas por um grande número de pequenas tarefas que solicitam acessos de forma randômica ao disco. Neste tipo de aplicação, tanto a banda passante do canal de comunicação quanto a latência podem ser apenas moderados, enquanto o *throughput* tem que ser bastante elevado. O maior desafio na definição de um subsistema de E/S é justamente conseguir um desempenho satisfatório tanto para as aplicações gerais quanto para as científicas, simultaneamente.

4 - OPÇÕES PARA CONFIGURAÇÃO DE UM SUBSISTEMA DE E/S

Conforme visto anteriormente, os microprocessadores são algumas ordens de grandeza mais rápidos que os dispositivos de armazenamento. Este desequilíbrio fez surgir um problema de adequação da alta velocidade de processamento com sua necessidade de entrada e saída de informações, evitando-se, assim, que o subsistema de E/S represente um gargalo para o sistema computacional. A solução deste problema pode apresentar algumas variações dependendo do tipo de aplicação a que a máquina se propõe, e será o tema da discussão a seguir.

O subsistema de E/S é um dos elementos de um sistema computacional. É composto de um ou mais processadores de E/S (PES) que controlam e gerenciam todas as operações na memória secundária. Existem várias formas de se associar estes PES aos elementos processadores do sistema. Dependendo da forma de associação consegue-se um maior ou menor desempenho em função do tipo de aplicação.

A primeira idéia que surge na definição de um subsistema de E/S é a utilização de um computador comercial específico para controlar todas as operações em memória de massa. Constituiria uma unidade completamente independente, interligada através de um canal de alta velocidade ao sistema principal. Aparentemente simples, esta idéia esbarra em alguns problemas que podem torná-la pouco recomendável na prática. Primeiramente, existe a dificuldade de se realizar eventuais modificações numa máquina comercial, onde o *hardware* e o *firmware* não são suficientemente abertos para permitir uma adaptação que proporcione um desempenho ótimo na execução desta nova função. Associado a isto, existe o questionamento do próprio desempenho desta opção frente às necessidades de E/S de um computador de alto desempenho voltado para processamento científico. Uma simultaneidade de pedidos de E/S com grande quantidade de dados pode congestionar o canal de comunicação. Por último, a necessidade de uma alta taxa de transferência neste canal é o maior dos problemas. Até o advento da tecnologia ótica, os canais de comunicação seriam muito lentos para este propósito. Canais como *Ethernet*, tipicamente implementados em uma única via de comunicação, operam a 10 MBits por segundo, taxa muito aquém das necessidades. Existe uma expectativa de que, com o uso da tecnologia ótica, bandas passantes de até 100 MBits por segundo por via de comunicação possam ser fácil e rapidamente alcançadas. Desta forma, com a tecnologia disponível hoje, a utilização de

canais de comunicação que concentrem toda a comunicação de E/S representaria um grande gargalo para o sistema. Entretanto, pode se caracterizar como uma técnica promissora.

Em contrapartida, a outra idéia que surge é a configuração do subsistema de E/S acoplado diretamente aos elementos processadores do sistema. Primeiramente, qualquer que seja a forma desta configuraçao, o fato dela ser interna ao sistema derermina um compromisso forte com suas características. O subsistema deixa de ser completamente independente para moldar-se a uma arquitetura específica, seguindo padrões elétricos e mecânicos que dificilmente se adaptariam a uma outra situação.

Pode-se basicamente dividir em duas as formas de se configurar internamente um subsistema de E/S. A forma distribuida, onde pequenos processadores de E/S (PES) são associados de forma exclusiva a cada um ou a pequenos grupos de elementos processadores, e a forma concentrada, onde um grupo de PES, acessado de forma simétrica por todos os elementos processadores, controla os dispositivos de armazenamento. Cada uma das formas tem suas vantagens e desvantagens como será mostrado adiante. Geralmente, as vantagens de uma são as desvantagens da outra e vice-versa.

E/S Distribuida:

Neste tipo de configuração, o processamento de E/S do sistema computacional está distribuido pelos seus diversos processadores de E/S. Cada PES está associado de forma exclusiva a cada elemento processador ou a um conjunto deles (*cluster*). Esta associação se realiza através de um canal de comunicação de alta velocidade, permitindo a transferência de informações entre os elementos processadores e o PES associado de forma satisfatória às necessidades de entrada e saída de um computador de alto desempenho. A exclusividade na associação não significa que somente os processadores de um mesmo *cluster* podem acessar o PES associado a eles. Qualquer elemento processador pode enviar e receber informações de qualquer processador de E/S. Entretanto, para tal, as transferências deverão ser realizadas através de caminhos alternativos, que seguramente terão uma banda passante menor que a do canal de comunicação do PES associado. Isto somente faz com que as transferências se processem com um custo de tempo mais elevado. Dentre as arquiteturas de E/S que se configuram como distribuida pode-se citar a do BBN-Butterfly [BBN 85] e a do RP3 [PFISTE 85].

Vantagens:

- *THROUGHPUT.* A fragmentação do subsistema de E/S pela distribuição de seu processamento nos diversos processadores de E/S permite um paralelismo das operações de entrada e saída. Cada PES pode transferir informações para qualquer elemento processador de seu *cluster* simultaneamente com os demais. Assim, o *throughput* ou a capacidade de processamento do subsistema de E/S é a soma do *throughput* de cada processador de E/S. Isto permite ao subsistema atingir um desempenho muito elevado, difícil de ser conseguido isoladamente por um único PES.

- SIMPLICIDADE. Como cada processador de E/S está associado a um pequeno grupo de elementos processadores, a quantidade de pedidos de entrada e saída que partem destes elementos é bastante reduzida se comparada com todo o subsistema. Isto acaba refletindo numa maior simplicidade de cada PES. O *throughput* deve ser o suficiente para atender às necessidades de entrada e saída apenas dos elementos processadores do *cluster*. Analogamente, a banda

passante do canal de comunicação entre o PES e o *cluster* associado também pode ser menor, resultando em técnicas de projeto mais simples. Além do mais, um menor número de pedidos de E/S permite ao PES respondê-los mais rapidamente, reduzindo também a latência do subsistema de E/S.

- MODULARIDADE. A configuração distribuida também tem a vantagem de ser modular. Cada módulo corresponde a um processador de E/S. Nesta configuração o subsistema de E/S pode crescer a medida em que cresce o número de elementos processadores do sistema, bastando para isso adicionar PES. Entretanto, não há a obrigatoriedade de existir sempre um PES associado a cada *cluster* de elementos processadores. Esta característica é importante na fase de implementação do sistema, onde apenas parte do subsistema de E/S pode ser implementado, deixando o seu crescimento vinculado ao crescimento progressivo de todo o sistema e à necessidade de reforço da capacidade de processamento de E/S.

Desvantagens:

- ESPECIFICIDADE. A distribuição do processamento de E/S torna o subsistema mais apropriado às aplicações específicas, como processamento científico e de imagens, pois muitas destas aplicações podem ser paralelizáveis. Isto se dá pela própria característica do processamento deste tipo de aplicação. São várias tarefas que podem ser executadas em paralelo em diferentes *clusters*, buscando, de cada PES associado, as informações necessárias ao processamento e reescrevendo, posteriormente, os resultados. Para aplicações gerais o desempenho deste tipo de configuração é questionável, principalmente quanto à distribuição dos arquivos ao longo dos PES.

- COERÊNCIA DE INFORMAÇÕES. Nesta configuração, os dispositivos de armazenamento estão distribuidos sob o controle independente de cada PES. Desta forma, existe uma dificuldade em se manter a coerência das informações gravadas neles. Ou seja, evitar que ocorra multiplicidade de informações e, principalmente, que as versões desatualizadas sejam invalidadas em todo o subsistema sempre que houver a atualização de algum arquivo.

- TRÁFEGO. Dependendo do tipo de aplicação a que o subsistema de E/S está submetido, a distribuição do processamento pode resultar em muitas transferências de informações entre *clusters* de elementos processadores. O custo de tempo destas transferências é bem maior que das transferências intra *cluster*, pois devem ser realizadas pela rede de comunicação entre elementos processadores de diferentes *clusters*. O aumento deste tipo de transferência pode saturar esta rede de comunicação, prejudicando o tráfego natural de troca de informações entre elementos processadores que se realiza através dela.

E/S Concentrada:

Na configuração concentrada o subsistema de E/S é composto por um grupo de um ou mais processadores de E/S. Este grupo gerencia e executa os pedidos de entrada e saída de todos os elementos processadores do sistema. A comunicação entre o grupo de PES e os elementos processadores é feita através de canais de comunicação. Também pode ocorrer a formação de *clusters* de elementos processadores com objetivo de acessar o subsistema de E/S, ou seja, a existência de um único caminho de comunicação com o grupo de PES para cada *cluster*. Dentre as arquiteturas de E/S que se configuram como concentrada pode-se citar a do ES-8701 [PRADO 88].

Vantagens:

- GENERALIDADE. A concentração do processamento de E/S permite uma adaptação melhor do subsistema de E/S à uma gama maior de variedades de aplicações. Isto ocorre por já haver intrinsicamente um compartilhamento de toda a memória secundária entre os elementos processadores. Toda informação é acessada por qualquer elemento processador com o mesmo custo de tempo. Entretanto, esta é uma característica que deve ser analisada em função do propósito da máquina. Da mesma forma que permite um desempenho satisfatório numa gama maior de aplicações, pode se tornar ineficiente em muitas aplicações científicas onde há a possibilidade de se explorar o paralelismo.

- GERÊNCIA DE TRANSFERÊNCIAS. Existem algumas políticas de otimização de acesso ao disco com objetivo de minimização do tempo médio de resposta a um pedido de E/S. Usualmente, se os pedidos forem atendidos obedecendo a ordem de chegada, pode resultar em movimentações extensas da cabeça de leitura/escrita ao longo das trilhas. Existem basicamente duas alternativas. Na política *Shortest-Seek-Time-First* os pedidos são atendidos na ordem em que minimiza o tempo de movimentação da cabeça a partir de sua posição atual. Uma outra opção é o algoritmo SCAN, onde a cabeça de leitura/escrita e movimentada de uma extremidade do disco à outra, atendendo aos pedidos de E/S à medida em que as informações neles solicitadas se encontrem progressivamente nas trilhas avançadas. Quando se concentra o processamento, pode-se explorar melhor estas políticas. Tem-se uma visão mais global dos pedidos de E/S e, portanto, é possível ordená-los de modo a gerenciar melhor as transferências.

Desvantagens:

- COMPLEXIDADE. A complexidade é a maior desvantagem desta configuração. A existência de um único grupo de processadores de E/S, onde se concentra todo o processamento de entrada e saída, pode simplificar o Sistema Operacional mas pode também requerer complexidade de *hardware*. Dependendo do número de elementos processadores, a demanda por processamento de E/S pode ser grande, implicando na necessidade de um alto desempenho do grupo de PES. São várias as dificultades inerentes a isto, afim de se evitar um gargalo no sistema. Desde o excesso de canais de comunicaçao entre o grupo de PES e os elementos processadores até a preocupação com a latência e o *throughput*, que requer o uso de técnicas de projeto mais sofisticadas.

-PREVISIBILIDADE. Não sendo uma configuração modular, o subsistema de E/S não cresce junto com o número de elementos processadores. Desta forma, o PES carrega desde o início as características de sua configuração final, requerendo um projeto que prevê todas as expansões futuras.

5 - ARQUITETURA DE E/S DO MULTIPLUS

O Multiplus é um multiprocessador científico de alto desempenho com arquitetura modular e memória global compartilhada. A arquitetura é capaz de suportar até 256 nós de processamento baseados em microprocessador RISC de 32 *bits* com arquitetura SPARC. Além do microprocessador, cada nó de processamento possui um co-processador de ponto flutuante, 16 M*Bytes* de memória pertencentes ao espaço de endereçamento global, *cache* de instrução e

dado separados, com 64 KBytes cada um, e *hardware* de suporte à gerência de memória.

Na arquitetura do Multiplus, um *cluster* é formado por até oito nós de processamento interligados por barramentos de dado e de instrução separados, de 64 *bits* de largura. Os *clusters* se comunicam através de uma rede de interconexão multiestágio do tipo n-cubo invertido, implementada com chaves *cross-bar* 2x2 com *fifos* nas saídas [AUDE 90]. A Figura 3 ilustra simplificadamente a arquitetura do Multiplus.

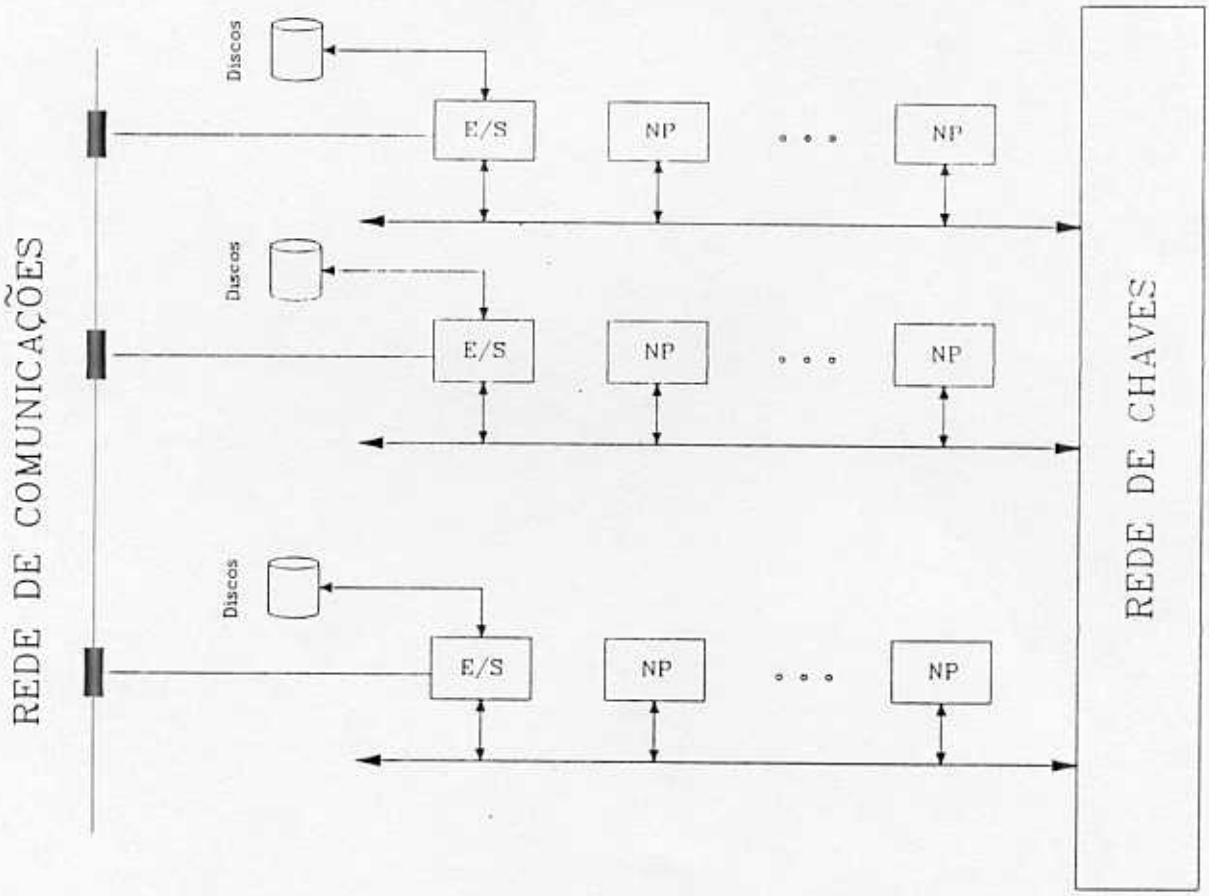

Figura 3: Diagrama simplificado do Multiplus.

Devido ao fato do Multiplus ser voltado para aplicações científicas, espera-se que suas operações de E/S sejam esparsas e extensas, devendo ser executadas no menor tempo possível. Isto caracteriza a necessidade de um canal de comunicação rápido e eficiente entre o subsistema de E/S e os elementos processadores. Entretanto, associar um canal de comunicação direto e de alta velocidade, partindo de cada elemento processador do Multiplus, ou até mesmo de cada *cluster* destes elementos até um único processador de E/S pode ser bastante custoso e inviável. Como o número total de *clusters* de elementos processadores é elevado, haveria uma grande quantidade de canais de comunicação. Além da dificuldade prática de interligação destes canais, haveria um subaproveitados devido à baixa

periodicidade das transferências. Descarta-se, desta forma uma configuração concentrada para o processamento de E/S do Multiplus.

Visualizando a arquitetura do Multiplus pode-se notar que ela própria sugere uma distribuição do processamento de E/S. Como são vários barramentos interligados por uma chave, é sugestivo associar um processador de E/S a cada barramento. Desta forma, o *cluster* de elementos processadores formado por cada barramento possui, através do próprio barramento, um canal de comunicaçao de alta velocidade entre os elementos processadores e o subsistema de E/S. Por outro lado, a própria finalidade a que o Multiplus se destina reforça esta configuração. Uma máquina paralela, como o nome já diz, é capaz de realizar várias operações em paralelo. E dentro destas operações também se encontram as de entrada e saída de dados e instruções. Este paralelismo é tanto mais explorado quanto mais distribuido for o processamento de E/S.

A comunicação entre os *clusters* de elementos processadores se dá através de uma rede de chaves multiestágio. Entretanto, o custo de tempo desta comunicação é alto frente às necessidades de E/S. Além do mais, sendo as transferências extensas, pode ocorrer a saturação da comunicação pela chave, em detrimento de sua *performance*. Assim, buscou-se um caminho alternativo para as transferências de E/S entre os *clusters*. Incorporou-se uma rede de comunicação ao subsistema de E/S, permitindo aos PES um caminho de comunicação entre si. Nota-se que para manter uma homogeneidade, esta comunicação é restrita ás transferências de entrada e saída. Uma outra característica interessante é que, com esta rede de comunicação, cada elemento processador solicita sempre serviços de E/S ao seu PES associado. Caso a informação solicitada esteja nos dispositivos de armazenamento controlados por este próprio PES, a transferência é realizada normalmente pelo barramento. Caso contrário, o PES solicita a informação ao PES apropriado, que a envia através da rede de comunicação e que, posteriormente, é transferida ao elemento processador via barramento. Entretanto, não é essencial que todo *cluster* possua um PES associado, podendo existir *clusters* só de elementos processadores. Neste caso, qualquer transferência de E/S é realizada obrigatoriamente pela rede de chaves.

Desta forma, o Multiplus se configura com um subsistema de E/S com duas forças de acoplamento aos elementos processadores. Cada *cluster* pode ter um PES fortemente acoplado aos seus elementos processadores, permitindo transferências com um custo de tempo bastante reduzido. Caso as informações de E/S não se encontrem no *cluster* do elemento processador solicitante, o acoplamento se enfraquece, aumentando o custo de tempo das transferências. Consegue-se, assim, explorar, a nível de *cluster*, as vantagens de uma configuração concentrada, embora o subsistema de E/S como um todo seja distribuido.

6 - O PROCESSADOR DE E/S

Um maior detalhamento do módulo Processador de E/S é mostrado no diagrama em blocos da Figura 4. O PES é dividido por dois barramentos. O barramento da CPU, onde também estão conectados a memória local e a memória de comandos/status, e o barramento interno, onde estão conectados as interfaces com os dispositivos de armazenamento, com a rede de comunicação e com os elementos processadores do Multiplus, além do DMA e da memória *cache*.

Esta divisão foi necessária para permitir ao DMA realizar transferências em rajadas (modo *burst*) sem interferir no processamento da CPU. Desta forma, a CPU, um microprocessador Motorola MC68020, é capaz de atender prontamente qualquer solicitação. E como ela controla todo o PES, isto permite um maior desempenho global, principalmente na gerência da memória de comandos/status, que é um recurso compartilhado, devendo ser liberado muito rapidamente. A memória local será, em princípio, de *4MBytes* com código de correção de erros. As transferências são realizadas por um DMA, e envolvem sempre a memória *cache*. Ou seja, as informações são lidas primeiramente para a *cache*, e, posteriormente, transferidas para o elemento processador. Isto permite que as transferências através do barramento de dados do Multiplus sejam feitas em modo *burst*, agilizando-as. A memória *cache* será de 16 *MBytes* também com código de correção de erros.

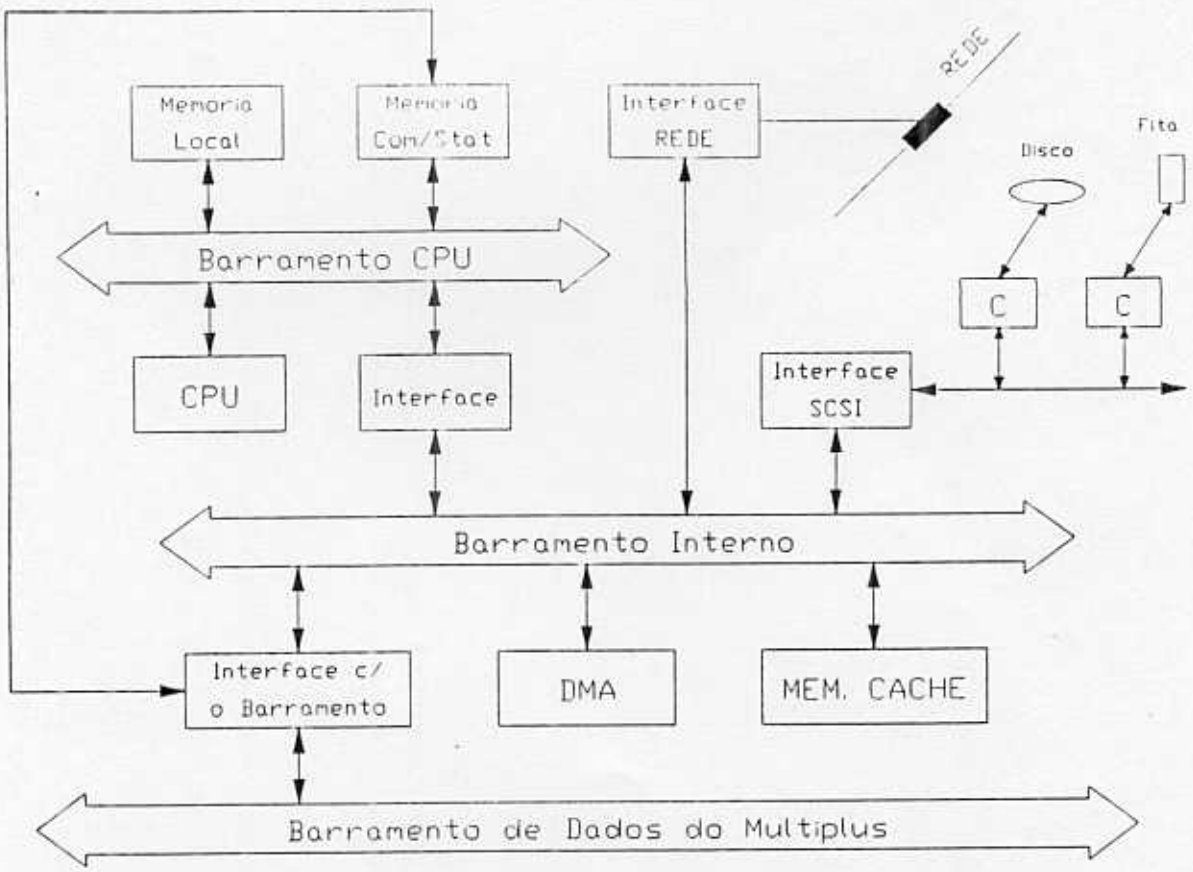

Figura 4: Diagrama em blocos do Proces. de E/S do Multiplus

Para avaliar se esta arquitetura é capaz de atender às necessidades de E/S do Multiplus pode-se partir de uma regra empírica bastante difundida na literatura: cada instrução por segundo demanda um *bit* de E/S por segundo. Cada PES deve suportar, no mínimo, a carga de E/S dos elementos processadores de seu *cluster*. Cada microprocessador SPARC do Multiplus é capaz de processar, efetivamente, 16 Mips a 25 MHz, e são no máximo oito SPARCs por *cluster*. Assim, o PES deve ser capaz de fornecer informações a uma taxa de 16 *MBytes/s*. Estas informações podem vir diretamente da memória *cache*, ou ainda através dos dispositivos de

armazenamento, dependendo da taxa de acerto na memória *cache*. Utilizando-se uma política *copy-back* em uma *cache* de 16 M*Bytes*, a taxa de acerto esperada é de 90% para as operações de leitura e de 65% para as de escrita de E/S. [FIGUEI 90]. Considerando ainda que 65% das operações de E/S são de leitura e apenas 35% são de escrita [FIGUEI 88], tem-se:

$$\text{Taxa média de acerto na } cache = (0.9 \times 0.65) + (0.65 \times 0.35) = 81\%$$

A banda passante da memória *cache* no caso do Multiplus é

$$B.P. = \frac{1}{\text{tempo de acesso}} \times (\text{Num. de } bytes/\text{palavra}) \times (\text{Num. de bancos})$$

supondo um tempo de acesso de 200ns, tem-se

$$B.P. = \frac{1}{200\ ns} \times 4 \times 1 \quad ====> \quad B.P. = 20\ MBytes/s$$

Desta forma, pode-se estimar a taxa de transferência necessária aos dispositivos de armazenamento:

$$\frac{1}{\text{taxa do PES}} = \frac{1}{\text{taxa do cache}} \times (\text{taxa de acerto na cache}) + \left(\frac{1}{\text{taxa do cache}} + \frac{1}{\text{taxa dos disp.}}\right) \times (\text{taxa de miss na cache})$$

$$\frac{1}{16\ MB/s} = \frac{1}{20\ MB/s} \times (0.81) + \left(\frac{1}{20\ MB/s} + \frac{1}{\text{taxa disp.}}\right) \times (0.19)$$

$$\frac{1}{16\ MB/s} = \frac{1}{20\ MB/s} + \frac{1}{\text{taxa disp.}} \times (0.19)$$

===> Taxa dos dispositivos = 15.2 M*Bytes*/s

Neste ponto depara-se com duas limitações. Primeiramente, para se explorar a banda passante da memória *cache*, é necessário que o DMA também seja capaz de transferir dados à uma taxa de 20 M*Bytes*/s. Entretanto, há um desinteresse por parte dos fabricantes destes dispositivos em desenvolverem versões que acompanhem a velocidade dos microprocessadores. O DMA utilizado, Motorola MC68450, é capaz de transferir a uma taxa máxima de apenas 5 M*Bytes*/s. A primeira versão do Projeto Multiplus não envolve o desenvolvimento de um DMA *custom* adequado, como, em geral, acontece em qualquer máquina de alto desempenho. A segunda limitação se dá na taxa de transferência dos dispositivos de armazenamento. Apesar de existirem discos magnéticos, tais como o CRAY DD-49 e CRAY DD-40, com taxa de transferência de até 9.6 M*Bytes*/s [PIEPER 89], a disponibilidade do Projeto Multiplus é de discos do tipo Winchester, cuja taxa de transferência é de 700 K*Bytes*/s. Assim, com estas limitações, pode-se estimar a taxa máxima de E/S que o PES é capaz de fornecer como sendo:

$$\frac{1}{\text{taxa do PES}} = \frac{1}{\text{taxa do DMA}} + \frac{1}{\text{taxa do WINCHESTER}} \times (\text{taxa de miss na cache})$$

$$= \frac{1}{5\ MB/s} + \frac{1}{0.7\ MB/s} \times (0.19)$$

====> Taxa máxima do PES = 2.12 M*Bytes*/s

Nota-se que esta taxa é cerca de 7 vezes menor que a requerida inicialmente. Entretanto, partiu-se de uma premissa de que 1 Mips demanda 1 M*Bit/s* de E/S. Analisando as arquiteturas de alto desempenho existentes, pode-se concluir que esta suposição é raramente cumprida. Mesmo porque, é uma suposição geral, aplicável a máquinas de propósito geral. Sendo o Multiplus uma máquina de propósito científico, espera-se que esta relação, entre a capacidade de processamento (Mips) e a necessidade de E/S, seja mais favorável, permitindo com que este PES satisfaça suas necessidades de E/S.

Existe uma alternativa, surgida nos últimos anos, de se organizar um conjunto de discos magnéticos afim de se obter uma maior taxa de transferência. Uma técnica chamada *stripping* [PATTER 88] propõe a fragmentação dos arquivos para armazenamento em uma matriz de discos, *Disk-Array*, de forma a cada fragmento ser armazenados em discos diferentes. Isto permite um paralelismo na leitura ou escrita de um bloco de informações, resultando, além da redução do tempo de *seek* e de latência rotacional, um aumento da taxa de transferência do conjunto de dispositivos de armazenamento. Há uma expectativa de se conseguir, através de *Disk-Array*, um subsistema de E/S que se adapte satisfatoriamente tanto às aplicações gerais quanto às científicas. Entretanto, devido a sua complexidade, é tema de estudo para versões futuras do Processador de E/S do Multiplus.

7 - ESTÁGIO ATUAL E PERSPECTIVA FUTURA

O subsistema de E/S de memória de massa do Multiplus se configura como distribuido. Seus dispositivos de armazenamento se distruibuem ao longo dos vários *clusters* de elementos processadores. Acredita-se que esta arquitetura de E/S explora melhor o paralelismo que é próprio das aplicações científicas, permitindo ao Multiplus um maior desempenho neste tipo de aplicação. Apesar das limitações impostas, espera-se que o subsistema de E/S satisfaça às necessidades de E/S do Multiplus.

Devido à sua simplicidade, e por não representar um gargalo para o sistema computacional, não foi mencionado neste trabalho os problemas encontrados na definição do Processador de E/S do Multiplus orientado à caracter. Este processador compõe, com o Processador de E/S de memória de massa, um par de processadores que, associados aos *clusters*, formam todo o subsistema de E/S do Multiplus.

No estágio atual do projeto deste subsistema, está se definindo o diagrama lógico dos Processadores de E/S. Há uma perspectiva de que, em meados do próximo ano, um protótipo do Multiplus esteja funcionando, juntamente com o subsistema de E/S.

8 - AGRADECIMENTOS

Os autores agradecem ao CNPq e a FINEP o apoio ao desenvolvimento deste projeto, bem como ao Eng. Norival Ribeiro Figueira pelas sugestões fornecidas na definição deste subsistema de E/S.

9 - BIBLIOGRAFIA

[HWANG 87] HWANG, Kai. Advanced Parallel Processing with Supercomputer Architectures. Proceeding of IEEE vol. 75, no. 10, october 1987.

[WILSON 87] WILSON, Ron. Designers Rescue Supercomputers from I/O Bottleneck. Computer Design, october 1, p.61-71, 1987.

[SWAN 87] SWAN, R. J. Cm* - A Modular, Multi-microprocessor. National Computer Conference, Montrale, New Jersey, AFIPS Press v.46 p. 637-663, 1987.

[GOTTLI 83] GOTTLIEB, Allan. The NYU Ultracomputer-Designing an MIMD Share Memory Parallel Computer. IEEE Transactions on Computers, vol. c-32, no.2, February, 1983.

[PIEPER 89] PIEPER, John S. Parallel I/O Systems for Multicomputers. School of Computer Science, Carnegie Mellon University, Pittsburgh, CMU-CS-89-143, 1989.

[CONNOR 87] CONNOR, Gary. Seial Data Races at Parallel Rates for the Best of Both Worlds. Eletronic Design, v.35, no.2, p.79-83, january 1987.

[MOKHOF 87] MOKHOFF, Nicolas. Five-chip Token-passing Set Operates LANS at 100 MBits/s. Eletronic Design, v.35. no.21, p.45-50, september 1987.

[PATTER 88] PATTERSON, D.A. A Case for Redundant Arrays of Inexpensive Disks (RAID). A.C.M. SIGMOD Conference, Chicago, IL, p.109-116, May, 1988.

[PRADO 88] PRADO, Cláudio Almeida. Projeto de um Subsistema de Memória de Massa Para Um Computador de Arquitetura Paralela. II-Simpósio Brasileiro de Arquitetura de Computadores, Águas de Lindóia, SP, pág 5.A.4.1-5.A.4.6, Setembro, 1988.

[FIGUEI 90] FIGUEIRA, Norival Ribeiro. Avaliação de Algoritmos Para Cache de Disco. Anais do XXIII-Congresso Nacional de Informatica, Rio de Janeiro, Setembro, 1990.

[FIGUEI 88] FIGUEIRA, Norival Ribeiro. Cache de Disco: Arquiteturas e Algoritmos. Anais do XXI-Congresso Nacioanal de Informática, Rio de Janeiro, vol.2, p.863-869, Agosto, 1988.

[BBN 85] BBN Laboratories Incorporated. Butterffly (TM) Parallel Processor Overview. June, 1985.

[PFISTE 85] PFISTER, G. F. The Architecture of the IBM Research Parallel Processor Prototype (RP3). Research Report, IBM T. J. Watson Research Laboratory, N.Y., June, 1985.

[AUDE 90] AUDE, J. S. MULTIPLUS: Um Multiprocessador de Alto Desempenho Anais do X-Congresso da Sociedade Brasileira de Computação, Vitória, E.S., p.93-105, Julho, 1990.